

# A novel method for transient detection in high-cadence optical surveys

## Its application for a systematic search for novae in M31

Monika D. Soraisam<sup>\*1</sup>, Marat Gilfanov<sup>1,2</sup>, Thomas Kupfer<sup>3</sup>, Frank Masci<sup>4</sup>, Allen W. Shafter<sup>5</sup>, Thomas A. Prince<sup>3</sup>, Shrinivas R. Kulkarni<sup>3</sup>, Eran O. Ofek<sup>6</sup>, and Eric Bellm<sup>3</sup>

<sup>1</sup> Max Planck Institute for Astrophysics, Karl-Schwarzschild-Str. 1, 85748 Garching, Germany

<sup>2</sup> Space Research Institute, Russian Academy of Sciences, Profsoyuznaya 84/32, 117997 Moscow, Russia

<sup>3</sup> Division of Physics, Mathematics, and Astronomy, California Institute of Technology, Pasadena, CA 91125, USA

<sup>4</sup> Infrared Processing and Analysis Center, California Institute of Technology, Pasadena, CA 91125, USA

<sup>5</sup> Department of Astronomy, San Diego State University, San Diego, CA 92182, USA

<sup>6</sup> Benoziyo Center for Astrophysics, Weizmann Institute of Science, 76100 Rehovot, Israel

Received date / Accepted date

### ABSTRACT

**Context.** In the present era of large-scale surveys in the time domain, the processing of the data, from procurement up to the detection of sources, is generally automated. One of the main challenges in the astrophysical analysis of their output is contamination by artifacts, especially in the regions of high surface brightness of unresolved emission.

**Aims.** We present a novel method for identifying candidates for variable and transient sources from the outputs of optical time-domain surveys' data pipelines. We use the method to conduct a systematic search for novae in the intermediate Palomar Transient Factory (iPTF) observations of the bulge part of M31 during the second half of 2013.

**Methods.** We demonstrate that a significant fraction of artifacts produced by the iPTF pipeline form a locally uniform background of false detections approximately obeying Poissonian statistics, whereas genuine variable and transient sources as well as artifacts associated with bright stars result in clusters of detections, whose spread is determined by the source localization accuracy. This makes the problem analogous to source detection on images produced by grazing incidence X-ray telescopes, enabling one to utilize the arsenal of powerful tools developed in X-ray astronomy. In particular, we use a wavelet-based source detection algorithm from the *Chandra* data analysis package CIAO.

**Results.** Starting from  $\sim 2.5 \cdot 10^5$  raw detections made by the iPTF data pipeline, we obtain  $\approx 4000$  unique source candidates. Cross-matching these candidates with the source-catalog of a deep reference image of the same field, we find counterparts for  $\sim 90\%$  of the candidates. These sources are either artifacts due to imperfect PSF matching or genuine variable sources. The remaining  $\sim 400$  detections are transient sources. We identify novae among these candidates by applying selection cuts to their lightcurves based on the expected properties of novae. Thus, we recovered all 12 known novae (not counting one that erupted toward the end of the survey) registered during the time span of the survey and discovered three nova candidates. Our method is generic and can be applied for mining any target out of the artifacts in optical time-domain data. As it is fully automated, its incompleteness can be accurately computed and corrected for.

**Key words.** Methods: data analysis – surveys – Novae, cataclysmic variables – galaxies: individual: M31.

## 1. Introduction

In time-domain astronomy, the difference imaging (DI) technique (Tomaney & Crotts 1996; Alard & Lupton 1998; Alard 2000; Alcock et al. 1999; Wozniak 2000; Bond et al. 2001; Gal-Yam et al. 2008) provides an efficient way to detect flux transients. In particular, as compared to the former conventional method of catalog cross-matching, DI provides a more effective strategy for finding variable and transient sources in crowded stellar fields. Such fields, as difficult as they are to analyze, hold equally rich opportunities for astrophysical studies. DI has thus become the staple choice as is evidenced by its implementation in the majority of the data analysis pipelines of modern time-

domain surveys, e.g., the intermediate Palomar Transient Factory (iPTF) survey (see Masci et al. 2016).

The basic principle of DI is to match the point-spread function (PSF) and the background, both of which often vary spatially, between a reference image and an input/science image. In the commonly used approach (e.g., Alard & Lupton 1998), this match is accomplished by convolving one of the images, generally the reference image with better seeing and higher signal-to-noise ratio, with a kernel that minimizes the difference between the convolved reference image and the science image. The determination of this kernel forms the most important step in such form of DI technique. An alternative and statistically more rigorous approach to implementation of DI for transient detection was recently proposed by Zackay et al. (2016). DI in its traditional formulation is implemented in many automated pipelines, particularly for large-scale surveys. Its execution, however, in

\* Present address: NOAO, Tucson, AZ, USA  
e-mail: monikas@mpa-garching.mpg.de

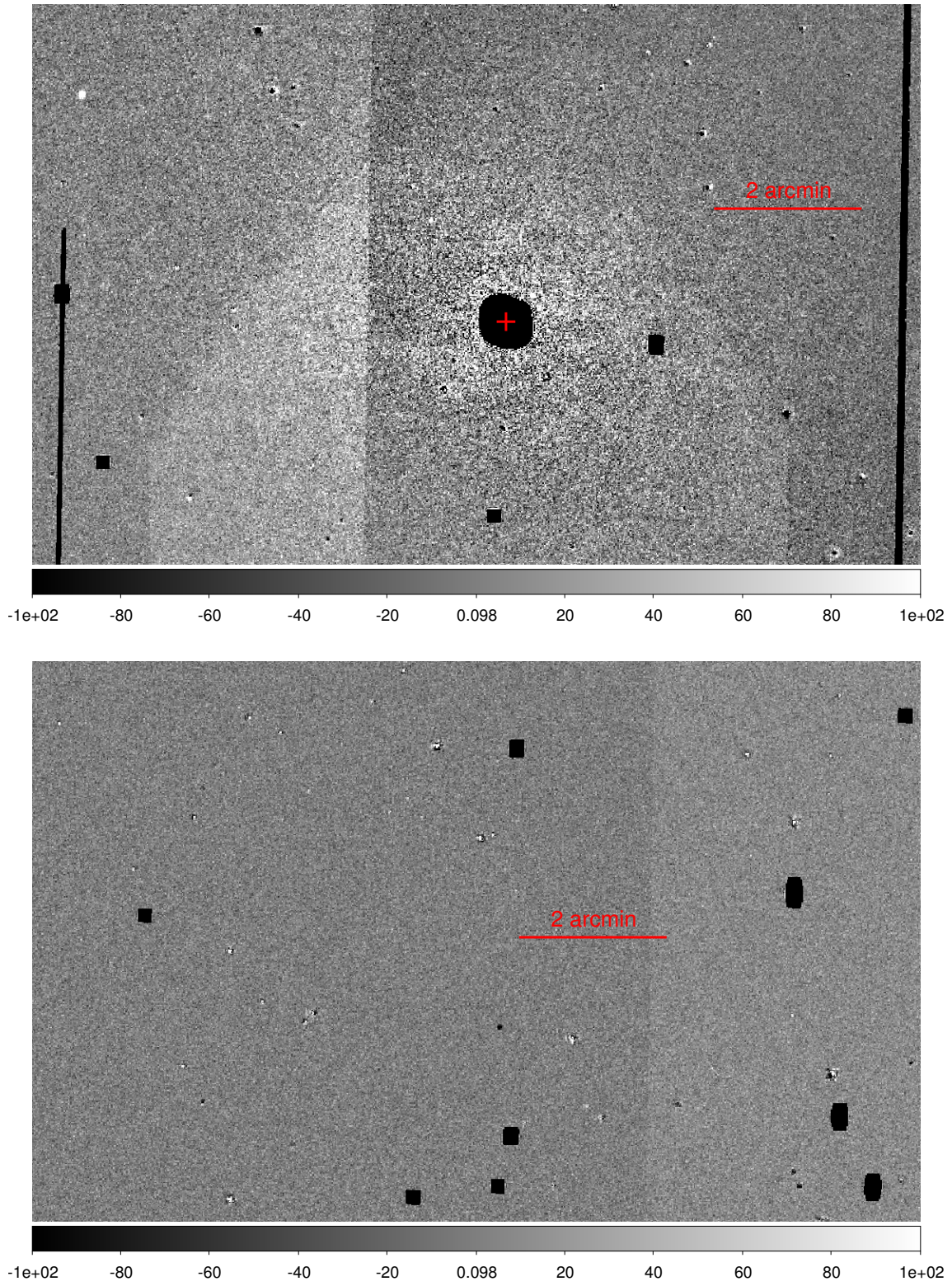


Fig. 1: *Upper panel*: A section of a typical M31 difference image from the iPTF pipeline containing the bulge. *Lower panel*: The outer part of the same difference image, about half a degree northward of the center of M31. The masked areas are assigned large negative values. North is upward and east is left. The center of M31 is marked by the red cross in the upper panel. The vertical sharp edges seen in both panels arise from the boundary of blocks that have been used to match the background before the image subtraction in the iPTF DI pipeline.

almost all cases is subject to too many false positives/artifacts compromising the quality of the resulting difference image, especially for the automated case. There are several reasons which may make the difference images susceptible to artifacts — the inherent instability of the mathematical process used (see Zackay et al. 2016 for details on this), registration errors, improper PSF and/or background matching, image edges, cosmic rays, etc. Consequently, the catalogs of sources detected in the difference images generally become contaminated, in fact dominated, by artifacts (see for example Bloom et al. 2012).

Furthermore, when dealing with a field containing a background that is spatially varying and bright, such as the bulge of M31, the DI quality tends to deteriorate further (Kerins et al. 2010). To illustrate this, we zoom in on two parts of a typical difference image of M31 generated by the iPTF DI pipeline (Masci et al. 2016), the bulge part and an outer part about half a degree northward of the center of M31, as shown in the upper and lower panels of Fig. 1, respectively. It is clearly evident comparing the two sections of the same difference image, that the residuals in the bulge part have much higher amplitudes than in the outer part. As such the bulge part is even more prone to artifacts. Thus, the artifact-contamination in the difference image catalogs for such fields worsens toward the bulge.

In this paper, using iPTF M31 observations as an example, we present a novel method to efficiently identify candidates for variable and transient sources in an artifact-dominated image and recover candidates even from the bulge, where the DI quality is low. We then make use of the method to conduct a systematic search for novae in M31. Although we describe our procedure using iPTF observations, the issues discussed above are generic and therefore our method is equally applicable to other surveys and other fields. The method involves mapping the sources detected from difference images onto a blank image of the corresponding field, thereby storing their spatial and recurrence information. We call the resulting image *spatial recurrence image*. In this image, the candidates for variable and transient sources appear as clusters of points over a background made up primarily of systematic artifacts, which are random and approximately follow Poisson distributions. In this latter aspect, the image becomes analogous to an X-ray image. Thus, exploiting the approximately Poissonian nature of the background, we make use of image analysis tools for source detection developed in X-ray astronomy, to obtain a list of unique candidates for variable and transient sources, significantly less contaminated by artifacts. In particular, we use the wavelet-based tool WAVDETECT (Freeman et al. 2002) from the *Chandra* Interactive Analysis of Observations (CIAO; Fruscione et al. 2006) software package. Once the candidates for variable and transient sources are obtained, we then follow the fairly standard practice of searching for novae — we construct the lightcurves for the candidates and then apply cuts based on expected properties of novae.

The paper is organized as follows. In Sect. 2, we describe the iPTF M31 data used in our analysis, and in Sect. 3, we present the spatial recurrence image of the sources detected in the difference images of the M31 observations by the iPTF pipeline. In Sect. 4, we describe the method we have developed to obtain the candidates for variable and transient sources from outputs of data pipelines of time-domain surveys. The procedure for searching novae amongst these candidates is outlined in Sect. 5. We show the results in Sect. 6 and end with their discussion and a summary in Sect. 7.

## 2. iPTF M31 data

The Palomar Transient Factory (PTF) survey (Law et al. 2009; Rau et al. 2009; Ofek et al. 2012) was succeeded by iPTF in 2013. Both share the same hardware and pipelines but some upgrades were implemented for the latter. The survey covers  $\approx 7.26 \text{ deg}^2$  on the sky with the camera mounted on the 1.2 m Samuel Oschin Telescope at the Palomar Observatory. The observations are conducted primarily in the *R* band, with some also made in the *g* band. The pixel scale of the detector is  $1.01''$  and the typical seeing is  $\approx 2''$ . With the nominal exposure time of 60 seconds, the  $5\sigma$  limiting magnitude reaches  $m_R \approx 21$ . The M31 field analyzed here, covering  $\approx 1 \times 0.5 \text{ deg}^2$  and including the bulge, was observed by iPTF between September 2013 and January 2014, comprising 201 epochs (all taken in the *R* band).

DI was performed for these epochs by the iPTF DI pipeline, which among other things performs the image convolution for PSF-matching using a technique developed by the iPTF team (Masci et al. 2016). The pipeline outputs as primary products photometric tables containing information (celestial coordinates, differential fluxes, etc.) on the sources detected (the detection criterion being signal-to-noise ratio  $S/N \gtrsim 3.5$ ) in the difference images. Users are responsible for further characterization of bona fide sources and the subsequent selection of objects of interest. Given that these detections are not artifact-free and thus are to be further cleaned, we here refer to these detections as *raw detections*. The typical number of raw detections per epoch extracted from the difference images by the iPTF pipeline is  $\sim 1000$ , and the total for this season is  $\sim 2.5 \cdot 10^5$ . The latter value of course includes multiple detections from any single source.

## 3. Spatial recurrence map of raw detections

With the goal of identifying unique variable and transient sources, we construct a spatial recurrence map of all the  $\sim 2.5 \cdot 10^5$  raw detections from Sect. 2. To this end, we create a blank image of the same M31 field, sampling the pixels at the iPTF scale of  $1.01''$  (Law et al. 2009). We then plot all the raw detections in this image, marking their positions with crosses. The resulting image, zoomed in on the bulge and on the outer part, is shown in Fig. 2 upper and lower panels, respectively.

The raw detections from individual sources cluster at the respective positions of the sources on the spatial recurrence image. These clusters can be clearly seen in a sea of single raw detections in the outer part of the M31 field in Fig. 2 lower panel. These give the locations of the candidates for variable sources, as well as transient sources appearing during the observation season, in the M31 field. This procedure thus provides a means for identifying unique variable and transient source candidates in one go, without the need for cross-matching the catalogs of raw detections from the DI pipeline. The bulge is, however, swamped by detections (see Fig. 2 upper panel). This is the cumulative effect of the low-quality DI at the bulge and higher concentration of real sources (cf. Sect. 1). Particularly for large-scale surveys in the time domain, the bulge part of a galaxy is thus one of the main challenges, where the bulk of the candidates are lost using conventional methods of thresholding on the parameters of the raw detections or even via machine-learning algorithms. Our method presented in this paper improves on this limitation of recovering candidates from the central parts of galaxies.

To facilitate further analysis, we first convert the spatial recurrence image (Fig. 2) to a counts image, representing each raw detection as one count. In the resulting counts image, which we



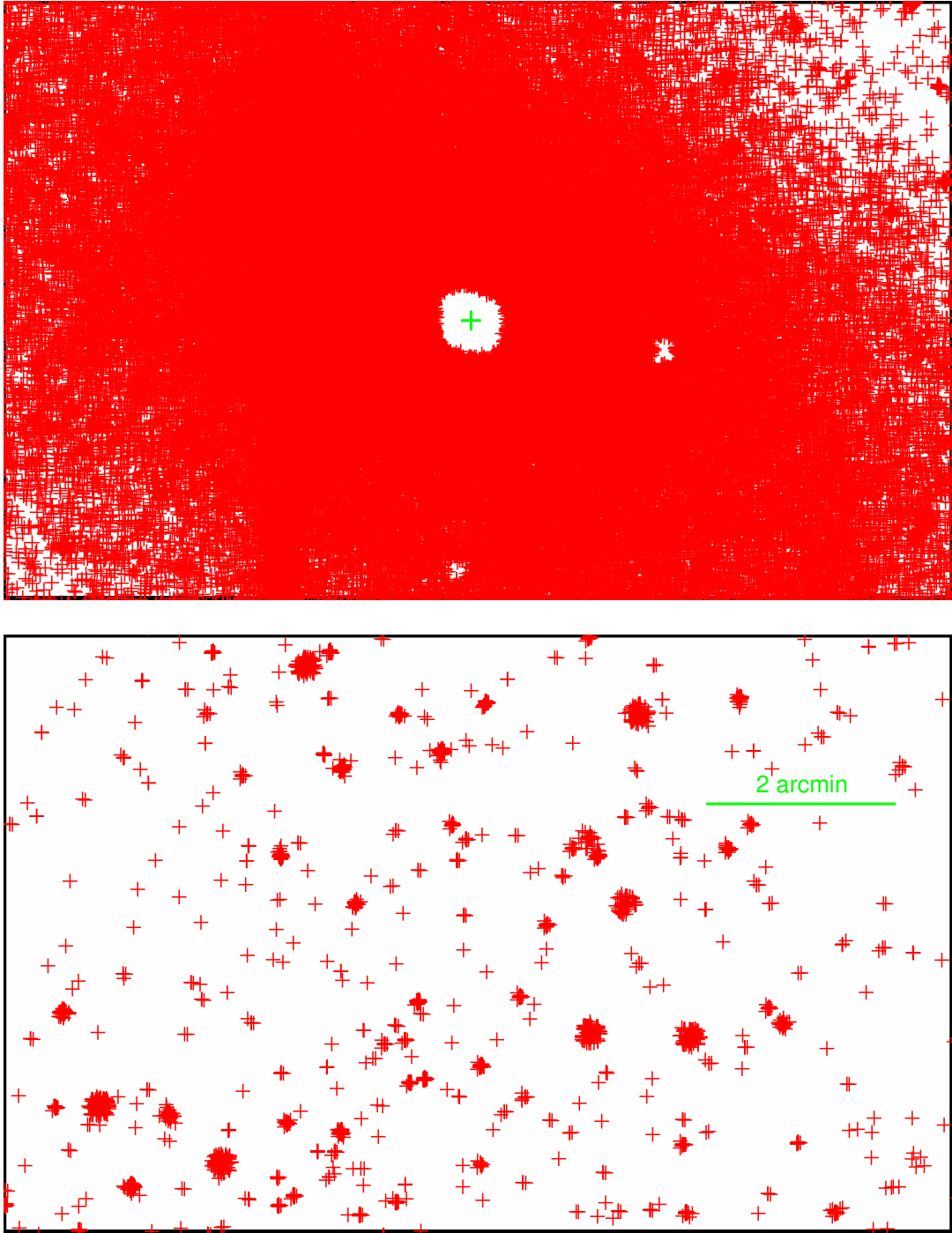


Fig. 2: *Upper panel*: The spatial recurrence image obtained by plotting the raw detections as red crosses (cf. Sects. 1, 3), zoomed in on the bulge of M31. The center of M31 is marked by the green cross. North is upward and east is left. Due to saturation close to the center of M31, there are no sources detected by the iPTF pipeline. *Lower panel*: The outer part of the spatial recurrence image, about half a degree northward of the center of M31.

call here the *hits image*, the pixel value equals the total number of occurrences of a raw detection within the given pixel. This hits image is shown in Fig. 3. Certainly some of the non-zero pixels in the image will be sources and the rest artifacts; the artifact contamination is expected to be more drastic in the bulge as can be inferred from the increased density of hits in that area (Fig. 3 upper panel).

#### 4. The method

In this section, we describe our procedure for identifying candidates for variable and transient sources in the M31 field analyzed. Only minimal information, namely the positions of all raw detections in an epoch, is required for the implementation of our method, and this is conveniently stored in the form of the

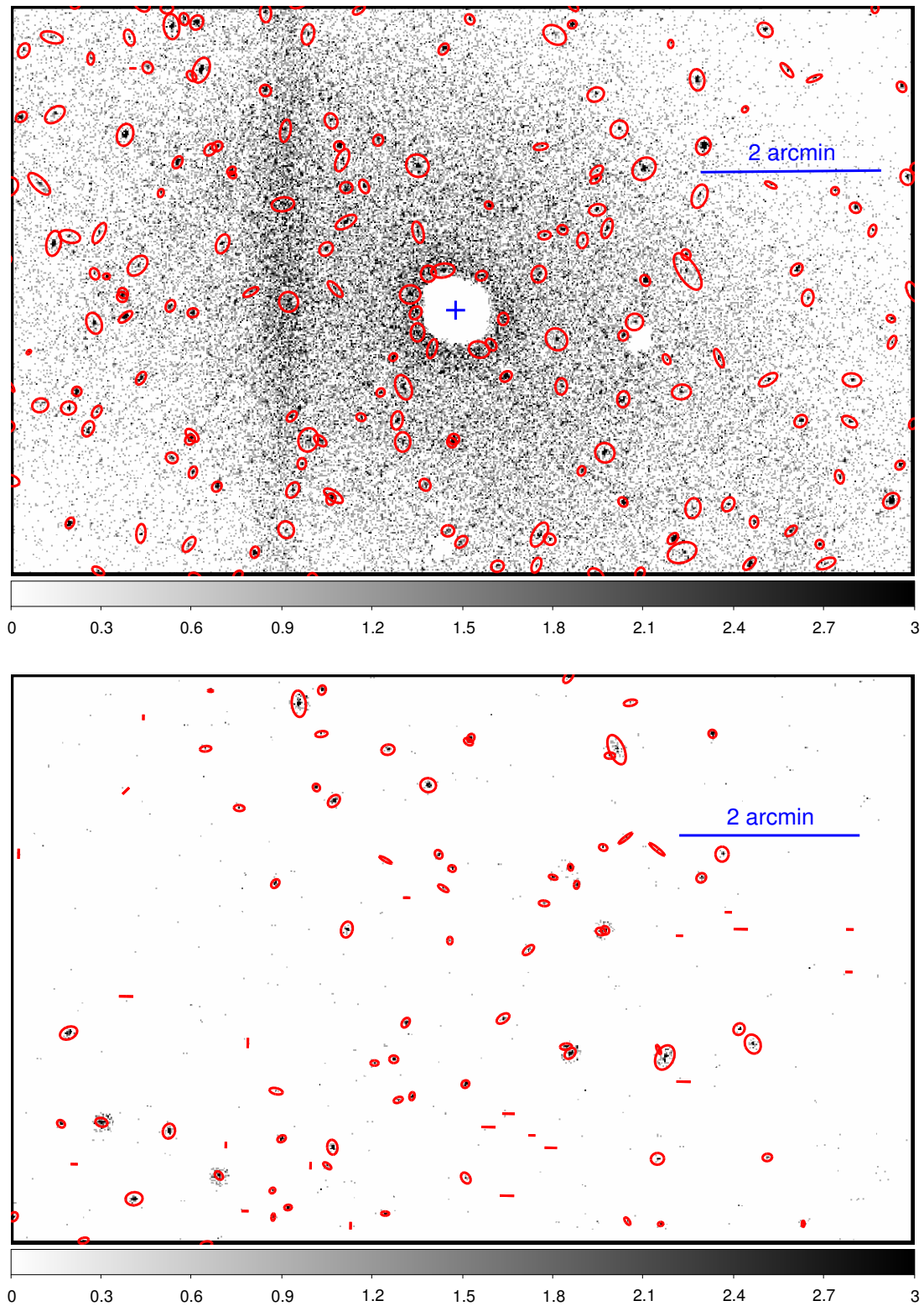


Fig. 3: *Upper panel:* The hits image obtained by counting the raw detections within each pixel (cf. Sect. 3), zoomed in on the bulge. The blue cross marks the center of M31 and the red ellipses show the location of sources found by WAVDETECT in this region. North is upward and east is left. As the area around the center is masked upstream in the iPTF pipeline due to saturation, the corresponding pixels in the hits image have zero values. The stripe that can be seen left of the center arises from the border between blocks used to match the background in the iPTF DI pipeline (cf. Fig. 1). *Lower panel:* The outer part of the hits image, half a degree northward of the M31 center, with the red ellipses showing the location of sources found by WAVDETECT. The shape of the elliptical region characterizes the shape of the distribution of counts within the source cell. Some of these regions output by WAVDETECT may have a very small length along one of the axes, such that they appear like lines; these regions are however only meant for visualization (see text).

hits image (Fig. 3). On this hits image, we perform a series of wavelet transformations using the CIAO tool **WAVDETECT**.

#### 4.1. Source detection using **WAVDETECT**

**WAVDETECT** is one of the most popular tools for X-ray source detection from the CIAO software package, based on wavelet analysis. In brief, it correlates the Mexican-hat wavelet functions with the input image. At every pixel  $(i, j)$ , the local background  $B_{i,j}$  is estimated using the negative annulus of the wavelet. Based on a user-defined significance value  $S$  (see Eq. 1), **WAVDETECT** computes a threshold correlation value ( $C_{i,j,0}$ ) for each pixel. It then identifies source pixels exhibiting correlation coefficients greater than their corresponding thresholds, and then these pixels are grouped into individual sources (see below) thus producing a final list of sources. Mathematically, **WAVDETECT** determines  $C_{i,j,0}$  as

$$S = \int_{C_{i,j,0}}^{\infty} p(C|B_{i,j})dC. \quad (1)$$

Here  $p(C|B_{i,j})$  is the probability of obtaining the correlation value  $C$  for the given background  $B_{i,j}$  at the pixel  $(i, j)$ , in the absence of a source in this pixel. This probability distribution has been determined from simulations (see Freeman et al. 2002 for details), and the results are already encoded within the software. If the correlation value  $C_{i,j}$  at pixel  $(i, j)$  is greater than  $C_{i,j,0}$ , then it passes through as a source pixel.

The correlation is performed at multiple wavelet scales defined by the user, where small (large) scales are sensitive to small (large) features. These scales  $s$  define the size of the wavelet; the wavelet crosses zero at  $\sqrt{2}s$  and extends effectively up to  $5s$ . The sources manifest themselves as islands (also called source cells) with some net counts above a background of nearly zero values when the image is smoothed with the positive part of the wavelet at the different scales and the background subtracted. The detected pixels are then assigned to their corresponding sources, and the final list of sources is made. The two parameters, namely the significance value/threshold and the wavelet scales, are thus the main inputs for source detection in **WAVDETECT**.

In processing the hits image via **WAVDETECT**, we adopt nine different scales ranging from 1 to 16 pixels, in multiplicative steps of  $\sqrt{2}$ . For the threshold, we choose a value of  $S = 10^{-6}$ . The M31 field analyzed here encompasses  $\sim 10^6$  pixels and therefore the above threshold implies an expectation value of  $\sim$  one false positive if the model for  $p(C|B_{i,j})$  is accurate (see Freeman et al. 2002).

Our **WAVDETECT** run found 3981 sources in the hits image. The upper and lower panels of Fig. 3 show the sources found in the bulge and half a degree northward of the center of M31, respectively, which are marked by the red ellipses. It is to be noted that the elliptical regions generated by **WAVDETECT** are only meant for visualization of the source locations, and do not affect the subsequent operation of **WAVDETECT** (CIAO; [http://cxc.harvard.edu/ciao/download/doc/detect\\_manual/wav\\_ref.html](http://cxc.harvard.edu/ciao/download/doc/detect_manual/wav_ref.html)).

The interpretation of the **WAVDETECT** result is that there are 3981 *different* locations from which multiple detections of the signal on the difference images occurred. Obviously, these locations are candidates for variable and transient sources in the M31 field obtained from the five-month-long iPTF season. This number may still be contaminated by artifacts due to statistical fluctuations that are not modeled accurately by the model for  $p(C|B_{i,j})$ .

However, the number is largely free of the background of single raw detections (cf. Sect. 3, Fig. 2 lower panel), as shown in Sect. 4.3. Of course, some of these single raw detections may be genuine sources. However, characterization of the nature of these single detections would require individual inspection of each of them, which is not feasible given the large number of artifacts produced by the iPTF data pipeline (but see Zackay et al. 2016 for an alternative approach to DI implementation).

#### 4.2. Applicability of **WAVDETECT** to the hits image

In the formulation of the **WAVDETECT** algorithm, the statistics of the background are tied to the probability distributions of the wavelet correlation values,  $p(C|B_{i,j})$  in Eq. (1). The algorithm by itself is quite versatile as it can be adapted for any given statistical characteristic of the background by computing the corresponding distribution of  $p(C|B_{i,j})$ . However, in the implementation of **WAVDETECT** within the CIAO software package, Poisson statistics for the background are assumed, as it is designed for X-ray data. In general for the algorithm, the statistics of the sources themselves do not come in the picture; what matters is only that they stand out above the Poissonian fluctuations of the background.

The second assumption of **WAVDETECT** is that the expectation value of the background is constant (Freeman et al. 2002). However, in real X-ray images, a constant background is definitely an idealization. In the execution of **WAVDETECT**, the background at a given pixel of the image is estimated from the region covered by the negative annulus of the wavelet centered at the pixel. Thus, the assumption of a constant expectation value for the background is effectively only made for a region of radius  $5s$  ( $s$  is the wavelet scale) around each pixel. Therefore, as long as there are no sharp variations in the background, the efficacy of **WAVDETECT** is not affected (Freeman et al. 2002).

When using **WAVDETECT** on the hits image, we are thus implicitly assuming the above conditions about its background — that it follows Poisson statistics and that the mean values do not vary sharply. In order to analyze how accurate these assumptions are, we use the background image (let us call it  $B'$ ) computed by **WAVDETECT**. In particular, we compare the pixel value distribution of the actual background of the hits image with that expected based on  $B'$ . To this end, we mask out circular areas, each of radius  $10''$  ( $\approx 10$  pixels), centered on the sources detected by **WAVDETECT**. For each of the remaining background pixels, we compute the expected distribution of counts in that pixel assuming Poisson distribution with the mean given by  $B'_{i,j}$ . We then sum these distributions and normalize the result to the total number of pixels in the sources-masked hits image. The resulting distribution is then compared with the actual pixel value distribution of the hits image with masked sources. As discussed in Sects. 1 and 3, the DI quality is systematically inferior in the bulge as compared to the disc (Fig. 1). We thus make the above comparison of the pixel value distributions separately for the bulge and disc. We define the bulge of the M31 field as an elliptical region of semi-major axis  $6'$  and axis ratio 0.47, centered on the M31 nucleus. The results are shown in the left and right panels of Fig. 4, respectively.

As can be seen from Fig 4, the background of the hits image does not strictly comply with Poissonian statistics, but the deviations are not stark. Moreover, even if the background were strictly Poissonian, one might expect that there could be some sources that barely stand out above the background. Those sources may not be detected above the threshold of  $S = 10^{-6}$

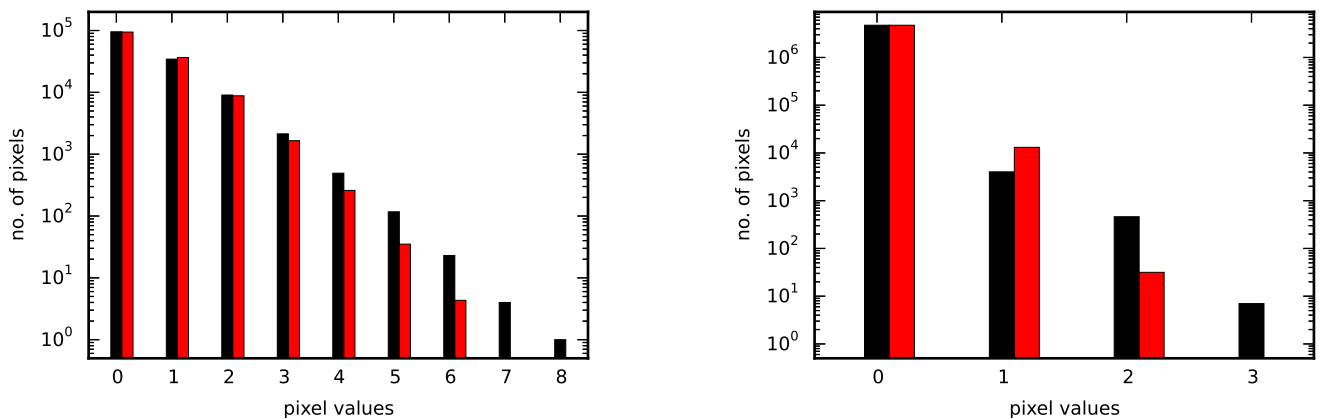


Fig. 4: Comparison of the background pixel value distribution of the hits image, shown in black, with the distribution expected for Poissonian background, shown in red, obtained using the WAVDETECT output  $B'$  (see Sect. 4.2), for the bulge (*left*) and for the disc (*right*). The red histograms have been shifted along the x-axis by pixel value=0.1 for clarity.

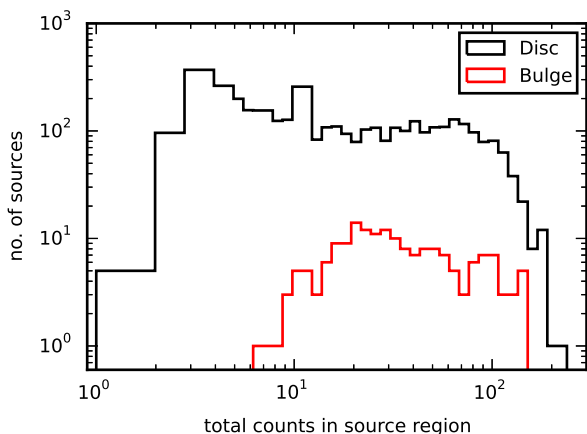


Fig. 5: Distribution of counts (i.e., the number of hits) in the source regions for the detections made by WAVDETECT in the bulge and the disc of the iPTF M31 field analyzed here. The five sources in the disc with one count passed through WAVDETECT due to their location in extensive empty areas.

that we have applied. Thus, these sources would not be masked and they would distort the observed histograms in Fig. 4. So even if the true background is exactly Poissonian and with a constant expectation value, the two histograms may not look precisely the same. In conclusion, the application of WAVDETECT on the hits image is justified to the accuracy sufficient for the purpose of this analysis.

#### 4.3. Characteristics of WAVDETECT detections in the time-domain context

It is quite evident from the range in the distribution of the background pixel values of the hits image (Fig. 4) that the detection threshold ( $C_{i,j,0}$  in Eq. 1) must be higher in the bulge than in the disc.

Besides the positions of the detected sources, WAVDETECT also includes in its output their *counts* — the sum of the pixel values in the corresponding source cells (see Sect. 4.1). In the context of the hits image, the source counts have the meaning of

the number of times a signal has been detected on the difference images from the given position. It is known that source counts computed by WAVDETECT are only crude estimates, since the tool is designed largely for detection and not for accurate photometry (CIAO). Of course, the former is our sole objective for use of the WAVDETECT tool. These crude source counts nevertheless allow us to gauge the difference in the sensitivity of WAVDETECT between the bulge and the disc of the M31 field, or in general between fields with sparse and crowded distributions of raw detections from the DI pipeline. The distributions of source counts in the bulge and disc of the hits image are shown in Fig. 5. As can be seen, the threshold number of hits for WAVDETECT detection is 7 in the bulge. In the disc, there are five sources detected with just one count; this arises due to these sources being located in extensive void areas, and this is a known feature of WAVDETECT (CIAO). However, they represent only 0.1% of the total number of detected sources (5 out of 3981), and are thus insignificant. Effectively, the threshold count in the disc is 2. The above discussion implies that, in the low background regions, WAVDETECT interprets source cells with two hits as sources; in the bulge region (Figs. 2 upper panel and 3), due to the high density of artifacts, at least seven hits within a source cell are required for reliable detection. Note that these values are for the adopted significance of  $S = 10^{-6}$  in running WAVDETECT (Sect. 4.1).

As discussed in Sect. 4.1, the 3981 sources detected by WAVDETECT in the hits image may not all be genuine sources. For instance, false positives may arise due to imperfect PSF-matching occurring persistently at the same position, for example at the position of a bright star. There may also be WAVDETECT false detections in the regions of high artifact density, for example caused by the deviations of the background distribution from being Poissonian. The latter is particularly relevant for those WAVDETECT sources at the low-count end of the distribution in Fig. 5. To assess broadly the nature of possible artifact-contamination, in particular related to intrinsic factors such as imperfect PSF and/or background matching, we cross-match the WAVDETECT sources with the sources detected in a deep reference image (co-add of 24 images) of the same field of M31. Out of 3981, we obtain matches for 3525 WAVDETECT sources. A matching radius of  $r \approx 2''$  — the typical full width at half maximum (FWHM) of the iPTF point spread function — was used in the cross-matching. The magnitude (in the  $R$  band) distribution



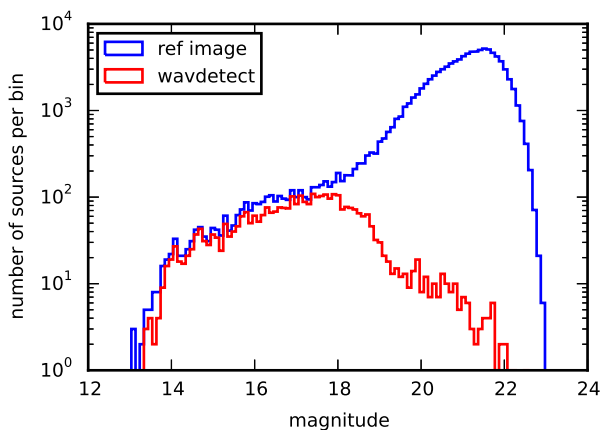


Fig. 6: Distribution of the magnitudes for the sources detected in the reference image and those WAVDETECT sources with counterparts in the reference image. The magnitudes of these WAVDETECT sources are measured from the reference image.

of the matched sources is shown in Fig. 6, along with that of all the sources in the reference image. Note that the magnitudes of the matched WAVDETECT sources are those measured in the reference image itself. Quite interestingly, the vast majority of bright sources in the reference image are detected in the difference images of the iPTF pipeline. Among these matches, some fraction may be genuinely variable stars, but some may be due to systematics in the iPTF DI pipeline caused by, for example, imperfect PSF and background matching procedure. Classification of artifacts from these matched sources is however beyond the scope of this work.

The total number of sources detected in the reference image down to  $m_R \approx 23$  is 99041. Thus, cross-matching the  $\sim 4000$  WAVDETECT sources with the reference image source catalog may lead to some number of false matches. To compute the expected number of false matches  $\langle N_f \rangle$ , we use Eq. (A.16) in Appendix A. Assuming a uniform density of reference image sources in the bulge and disc separately, we compute  $\langle N_f \rangle$  in the respective region. We then sum them and obtain the total number of expected false matches as  $\approx 60$ , which is about 2% of the matched sources.

The remaining 456 WAVDETECT sources are then candidates for transients, i.e., objects without counterparts in the deep reference image (different from variable sources in this context). The classification of these sources is an ongoing project and is beyond the scope of this paper. The factors that can cause artifact-contamination for this subset, if any, may include image edges and imperfect background matching. The latter, however, has to occur persistently in the same location on the sky. This may happen due to imperfect filtering of detector and optical artifacts when creating the reference image used for the subtraction, thus contaminating all the resulting difference images obtained with the given reference image.

In Table 1, we summarize the characteristics of the WAVDETECT detections. For transients, our method thus scales down the search sample by nearly three orders of magnitude from the raw detections of  $\sim 2.5 \cdot 10^5$  to a sample of a few  $10^2$  candidates. However, it is to be noted that the 60 false matches from above constitute about 12% of the total number of transient candidates. Depending on the task at hand, one may choose to minimize the number of candidates and consider only the 456

sources for characterization and follow-up studies or take a more conservative approach and take into consideration all 3981 candidates. For example, the number of candidates may be critical for the feasibility of follow-up spectroscopy, which is resource expensive. On the other hand, lightcurve filtering can be easily done for all 3981 candidates. Taking the conservative approach is also relevant for our nova search. For this case, besides the existence of false matches, there are two more reasons to consider all candidates (with and without matches in the reference image). Firstly, the reference image used for cross-matching has been constructed (via co-addition) using a subset of observations from the same data set used for searching the novae and therefore it can be contaminated by some novae. Secondly, there is no sharp cut-off for the quiescence magnitudes of novae; for example, the quiescence magnitudes of novae in M31 measured by Williams et al. (2014) reach values of  $m_R \sim 22$ .

## 5. A systematic search for novae in the iPTF M31 observations

With the set of unique candidates for variable and transient sources obtained via WAVDETECT (Sect. 4.1), it becomes feasible to search for any class of objects of interest. Hereupon we simply need to resort to the routine procedure of constructing lightcurves and making the selection based on the expected behavior of the objects sought. With an efficient algorithm, the few thousand candidates obtained in Sect. 4.1 can be reduced to a set of a few, optimized for the class of objects of interest. In particular, we make a systematic search for novae among the 3981 variable and transient source candidates found above.

Once the locations of the different candidate sources are identified, we proceed to construct their lightcurves using the whole set of *difference images* per candidate. We perform forced photometry at these positions, i.e., we measure the fluxes of the candidates at all epochs irrespective of whether or not they are detected in a given epoch by the iPTF DI pipeline. As we are measuring fluxes on difference images, which are relatively sparsely populated and stellar blending is not expected, we have opted for simple aperture photometry. This is executed using the PHOT package of DAOPHOT (Stetson 1987) and with curve of growth correction using DAOGROW (Stetson 1990). The resulting flux measurements thus represent the differential fluxes of the variable and transient source candidates. Note that the measured fluxes are in detector units, i.e., DN.

Four examples of the resulting lightcurves are shown in Fig. 7. As can be seen in the figure, three of the lightcurves exhibit clear and distinct features of variability; in fact, one is the lightcurve of a nova (Fig. 7 top right panel), as we will confirm in Sect. 6. The remaining fourth lightcurve is almost flat and featureless. From visual inspection, we verified this latter candidate to be an artifact resulting from imperfect PSF-matching in the DI. Further, some outlier data points can be clearly seen, for example in the bottom right panel of Fig. 7. These spurious flux measurements also arise from low-quality difference images.

The lightcurves of all the variable and transient source candidates found by WAVDETECT are then filtered for novae, as described in the following subsection.

### 5.1. Lightcurve filtering for novae

We expect a nova lightcurve to exhibit a contiguous or pseudo-contiguous (see below) section of excess brightness marking the eruption, while being consistent with the quiescent background



Table 1: Summary of WAVDETECT detections

Raw detections from iPTF pipeline (not necessarily unique sources)	254765
Number of unique sources obtained from WAVDETECT	3981
Total number of raw detections encompassed by the WAVDETECT sources	120287
Number of WAVDETECT sources with counterparts in the reference image	3525
Expected number of false matches in the above	60
Number of transient candidates	456

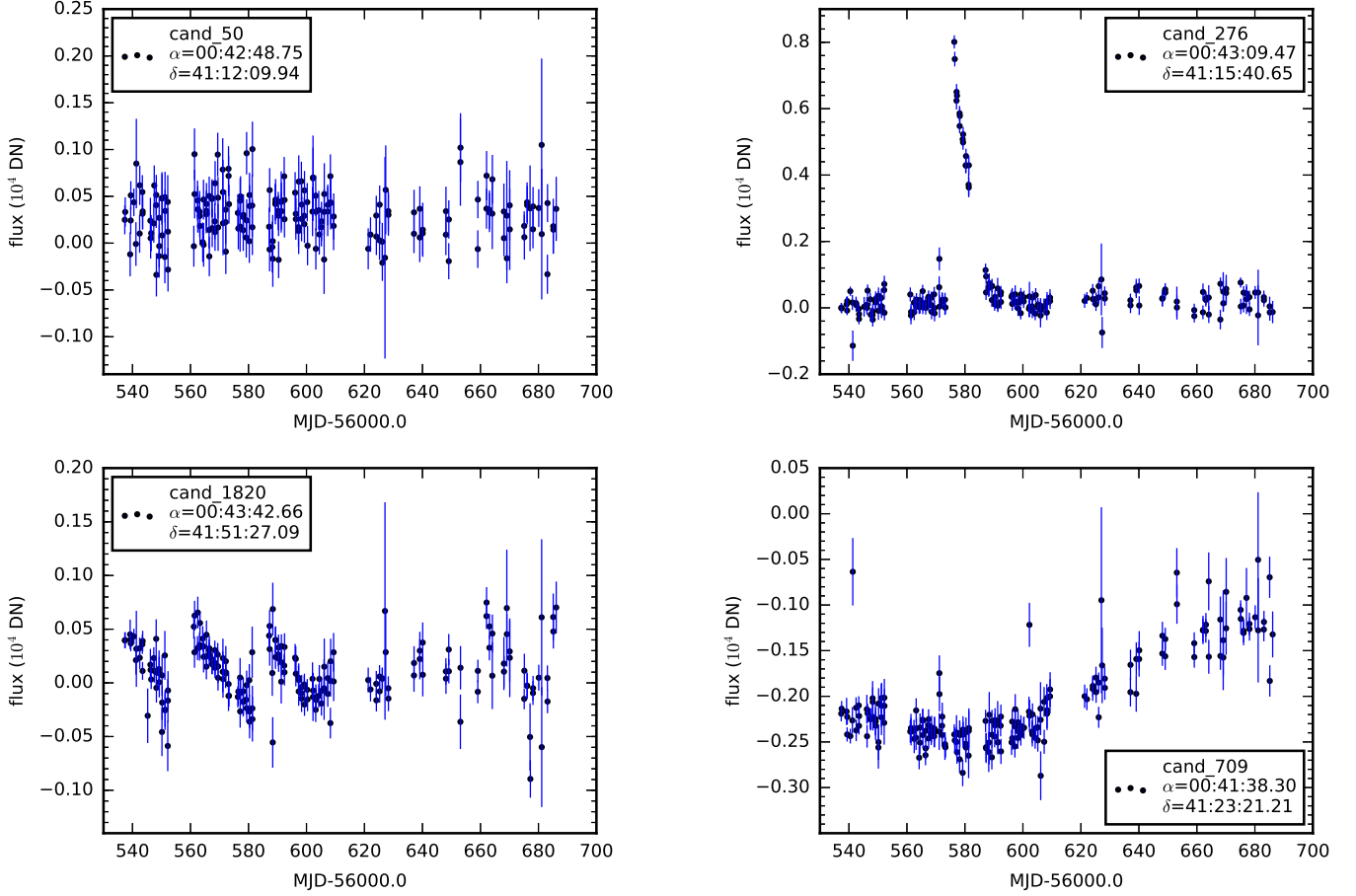


Fig. 7: Examples of lightcurves of four candidates identified by WAVDETECT, constructed via forced aperture photometry on the difference images. The lightcurve in the top left panel is that of an artifact resulting from imperfect PSF-matching in the DI, while the one in the right panel is that of a nova (cf. Sect. 6). The two lightcurves in the bottom are that of genuine variable sources. For the source in the bottom right panel, all the measured fluxes have negative values. This occurs because the reference image used in the difference imaging is not an unbiased average of the random phases of the lightcurve.

before and probably after the eruption. We build our lightcurve-filtering algorithm on this premise. The steps implemented are the following:

1. A baseline is determined iteratively by computing the mean of the 2-sigma clipped lightcurve using all the data points. The clipping is carried out both above and below the mean, using the rms deviation about the mean. In case there are fewer than 10 points remaining before convergence of the clipping process (for example this might happen in periodic

variables), we use all the data points in computing the mean. The final mean and the rms deviation are then used as the baseline and  $\sigma$ , respectively, in the following steps.

2. All the data points in the lightcurve that are more than  $4\sigma$  above the baseline are identified. Let us call this set of points  $S_p$  and the number of points in it  $n_p$ . It does not matter here whether these points are contiguous or not.
3. All the data points in the lightcurve that are more than  $4\sigma$  below the baseline are also identified. Let us call this set  $S_n$  and its number of points  $n_n$ .

4. We require that the nova candidate lightcurves have  $n_p \geq 4$  and  $n_n \leq \max\{3, 0.30 \times n_p\}$ . The latter condition helps in filtering out periodic variables.
5. We take the observation times (i.e., the modified Julian date MJD) corresponding to the first and last points of  $S_p$ ,  $t_{\min}$  and  $t_{\max}$ , respectively, to bracket the main outburst part of the nova. Let  $n$  denote the *total* number of data points in the lightcurve between  $t_{\min}$  and  $t_{\max}$ . If  $n_p \geq 0.70 \times n$ , then step 7 is executed, else step 6. The factor 0.7 is chosen somewhat arbitrarily through experimentation.
6. If  $n_p < 0.70 \times n$ , the following steps are implemented to address the possible presence of outlier flux measurements (see above) contaminating potential nova candidate lightcurves.
  - (a) In principle, for uniform temporal coverage, we could do the following. We can take the observation MJDs corresponding to the  $S_p$  points and compute an array of interval in time between consecutive  $S_p$  points. The spurious data points will have large values and will be outliers in such an array. The iPTF coverage however is not uniform, having gaps that could be even greater than a week. Nevertheless, the above method can still be implemented by working simply with array indices, as detailed below. We first make an array containing all the data points of the lightcurve, sorted in increasing order of the observation MJDs. Using the indices ( $i$ ) from this array corresponding to the  $S_p$  points, we create a new array, let us call it *index-interval array*, containing their forward differences (i.e.,  $i_{k+1} - i_k$ ). We can get rid of outlier data points in  $S_p$  by partitioning the index-interval array using some statistics of the array itself. We choose to use the median  $M$  and the rms deviation about the median  $\sigma_M$ , determined from the 3-sigma clipped index-interval array. The clipping is performed iteratively *above* the mean of the array, using the rms deviation about the mean. If convergence is not achieved by the time the number of elements in the clipped array reduces to 80% of the original number, we proceed with the immediately preceding clipped index-interval array. Using the so-determined  $M$  and  $\sigma_M$ , we partition the original index-interval array into segments  $L_i$  at positions where the array value is  $> \max\{2, M + 1.5\sigma_M\}$ .
  - (b) We take the longest segment  $\max(L_i)$  as the new  $S_p$ , with  $n_p$  denoting its number of elements. We then determine  $t_{\min}$ ,  $t_{\max}$  and  $n$  as in Step 5. If  $n_p \geq 0.70 \times n$ , step 7 is executed, else the candidate is discarded.
7. Finally, we apply a rise-time filter to the lightcurves that have passed through step 5 to eliminate long-period variable stars. In a nova, the initial rise takes at most 3 days; this is followed by a pre-maximum halt that lasts for a few hours to few days, and the final  $\sim 2$  mag rise to the peak takes some days to weeks (Warner 1995). Accordingly, we require the rise-time of the candidate lightcurve to be less than 2 weeks. The rise-time is defined as  $t_{\text{peak}} - t_{\min}$ , where  $t_{\text{peak}}$  is determined by fitting a broken power law to  $S_p$  with one breakpoint corresponding to the peak.

After passing the 3981 candidate lightcurves through the above selection algorithm, we obtain a set of 15 nova candidates, which are listed in Table 2. Estimates of the lightcurve peak value (obtained from Step 7 above) expressed in calibrated magnitude  $m_{R,\text{peak}}$  and decline times  $t_2$  are also included in the table. These  $t_2$  values have been obtained by linear interpolation of the data points; when fluctuations are present in the lightcurves, we give values of both the *first* ( $t_{2i}$ ) and *last* ( $t_{2f}$ ) times the lightcurve

falls by 2 mags from peak (or in linear scale when the flux has declined by a factor of  $10^{2/2.5}$  from the peak value). We determine the accuracy of the  $t_2$  estimates by employing a bootstrapping algorithm, whereby we resample a given lightcurve 1000 times by randomly selecting half of the data points and recomputing the  $t_2$  values each time. We then obtain the uncertainty by computing the standard deviation of the 1000  $t_2$  values<sup>1</sup>. It is to be noted that for candidates 1626 and 355, the flux corresponding to 2 mags from the peak is within  $1\sigma$  of the baseline already, and therefore for these two candidates, a better estimate of their  $t_2$  could be provided by  $t_{2i}$ . Furthermore, for candidates 1687 and 981, their  $t_2$  times cannot be estimated as their lightcurves have not yet declined by 2 mags from the peaks in the period analyzed here. Of course, slow novae, whose duration of eruptions is comparable to or much longer than the baseline of the observations presented here (five months) will be missed in our sample.

## 6. Results

The lightcurves of the final 15 nova candidates are shown in Fig. 8. The identification number (ID) of the candidates from Table 2 are included in the legends on the plots. From the M31 optical nova catalog maintained at MPE<sup>2</sup> (Pietsch et al. 2007), we are able to verify that 12 of them are in fact confirmed novae. The names of these candidates in the MPE catalog are also shown in Table 2.

During the five-month period between September 2013 and January 2014, when these data were taken, 13 novae (occurring in the field examined here) have been registered at the MPE catalog by different authors based on observations with different telescope facilities. We list the 13 novae in Table 2. We have recovered all except one, i.e. M31N 2014-01c. This nova occurred close to the center of M31 at the end of January 2014, which is near the end of the season analyzed here. The iPTF DI pipeline detected this nova only 5 times when it erupted. This value is, however, insufficient for its detection in the densely crowded bulge (cf. Fig. 3) by WAVDETECT, which requires at least 7 hits (see Sect. 4.3). With data extending beyond January 2014, this nova would most likely have been detected by WAVDETECT. For the remaining three nova candidates in our sample (1626, 1687 and 3182), candidate 1626 appears clearly a nova, but for candidates 1687 and 3182, we cannot firmly establish their nova nature from the present lightcurves since their variability is detected at the end of the observation period analyzed here. Still, candidate 1687 could be a fast faint nova similar to M31N 2014-01b (ID 981). In this work we have thus found one new nova and two nova candidates in M31 analyzing the five-month-long data set.

From the discussion above, the MPE nova catalog appears to be comparable with the iPTF nova sample obtained here. The former is mainly a result of work of the many amateur astronomers who made the discoveries of the individual novae. However, data from surveys are better-suited for statistical studies of the nova population, since surveys provide the advantage of uniform coverage in various aspects of the observations, for example detectors used, that allows a robust computation of the sample completeness (see also Sect. 7). The fact that the whole

<sup>1</sup> Due to the correlated nature of the time-series data, the bootstrapping algorithm does not strictly provide a formal one-sigma error bar. The quoted uncertainties should rather be interpreted as a rough indication of the precision of the  $t_2$  estimates.

<sup>2</sup> [http://www.mpe.mpg.de/~m31novae/opt/m31/M31\\_table.html](http://www.mpe.mpg.de/~m31novae/opt/m31/M31_table.html)

Table 2: iPTF M31 nova sample for the period between 09/2013 and 01/2014.

ID	RA <sup>a</sup>	DEC <sup>a</sup>	$m_{R,\text{peak}}$	$t_2(\text{days})^{b,c}$	Known nova name	Reference	Comments <sup>d</sup>
89	0:43:00.47	41:12:36.71	15.8	$10.3 \pm 0.3$	M31N 2013-09c	1	
234	0:43:13.55	41:14:47.69	16.4	$11.2 \pm 2.6$	M31N 2014-01a	2	spectroscopically confirmed; type Fe II
243	0:42:49.87	41:14:57.81	18.0	$28.0 \pm 4.9 - 44.6 \pm 4.8$	M31N 2013-12a	3	spectroscopically confirmed; type narrow-lined He/N
276	0:43:09.47	41:15:40.65	17.3	$10.4 \pm 0.4$	M31N 2013-10c	1	
319	0:43:04.90	41:16:31.13	14.8	$6.4 \pm 0.3$	M31N 2013-10h	4	spectroscopically confirmed; type Fe II
355	0:42:46.52	41:17:00.84	16.7	$24.2 \pm 3.5 - 83.0 \pm 7.2$	M31N 2013-10b	5	
376	0:42:42.35	41:17:18.86	16.0	$7.0 \pm 0.2$	M31N 2013-10e	1	
427	0:43:24.21	41:18:19.49	16.4	$19.6 \pm 0.5 - 21.6 \pm 0.9$	M31N 2013-10a	6	spectroscopically confirmed; type Fe II
430	0:42:51.66	41:18:14.71	16.7	$13.1 \pm 1.1 - 16.1 \pm 1.4$	M31N 2013-12b	7	spectroscopically confirmed; type hybrid
473	0:43:14.93	41:19:13.35	16.5	$12.1 \pm 0.6$	M31N 2013-09a	8	
601	0:43:24.90	41:21:22.38	17.4	$43.8 \pm 1.8 - 47.1 \pm 1.2$	M31N 2013-10g	9	spectroscopically confirmed; type Fe II
981	0:42:23.55	41:29:12.68	18.9	–	M31N 2014-01b	10	spectroscopically confirmed
1626	0:41:47.02	41:44:11.59	19.1	$28.6 \pm 8.6 - 110.8 \pm 4.7$	–	this work	possible nova
1687	0:43:50.44	41:46:12.36	19.3	–	–	this work	nova candidate (faint, fast)
3182	0:43:15.56	42:13:51.34	19.5	$3.0 \pm 1.3 - 8.9 \pm 2.3$	–	this work	nova candidate
–	0:42:34.94	41:14:56.30	–	–	M31N 2014-01c	10	spectroscopically confirmed; type Fe II. Erupted at the end of the iPTF season analyzed here

**Notes.** <sup>(a)</sup> ICRS coordinate system (J2000.0). <sup>(b)</sup> When there are fluctuations in the lightcurve, two values of  $t_2$  are given corresponding, respectively, to the first and last times the lightcurve falls by 2 mags from peak. <sup>(c)</sup> The errors on the  $t_2$  values are estimated by a bootstrapping algorithm (see text). <sup>(d)</sup> Information about the spectroscopic types for the novae here is included whenever available (for details, see for example, Williams 1992, Shafter et al. 2011).

**References.** (1) Hornoch et al. (2013); (2) Fabrika et al. (2014c); (3) Fabrika et al. (2014a); (4) Fabrika et al. (2013b); (5) Hornoch & Vrástil (2013); (6) Bigley et al. (2013); (7) Fabrika et al. (2014b); (8) Sturm et al. (2013); (9) Fabrika et al. (2013a); (10) Tang et al. (2014).

procedure involved in this search for novae is automated, makes it possible to quantitatively characterize any incompleteness in the sample. This is left for future work, following the analysis of the much larger full iPTF data set of M31.

## 7. Discussion and summary

We have presented a new and efficient method for tackling the menace of artifacts contaminating the genuine variable and transient sources detected by automated data pipelines of time-domain surveys. We have used a five-month-long string of iPTF observations of M31 covering its crowded bulge, with  $\sim 2.5 \cdot 10^5$  raw detections by the iPTF DI pipeline, to illustrate the methodology. Using all the pipeline detections, we created an image, termed the “hits-image”, each pixel of which records the number of times a signal from the given position was registered on the difference images (Sect. 3, Figs. 2, 3). We have shown that about half of the artifacts form a locally uniform (nearly) Poissonian background in the hits image (Fig. 4), with the remaining half being associated with bright stars. The latter, as well as genuinely variable and transient sources appear as clusters of detections, much like in an image produced by a grazing incidence X-ray telescope (cf. Fig. 2 lower panel). This analogy with an X-ray image allows us to import various well-established tools in X-ray astronomy without requiring any modification. Multi-scale wavelet analysis provides an efficient and convenient means to detect structures/sources in such images. We have thus chosen the popular wavelet-based tool WAVDETECT from the *Chandra*’s CIAO package (see Freeman et al. 2002) for the identification of the clusters of detections in the hits image. Running WAVDETECT on it, we obtained the unique candidates for variable and transient sources for the whole observation season, numbering  $\sim 4 \cdot 10^3$ .

Cross-matching the WAVDETECT “sources” with the source-catalog of a deep reference image of the same M31 field, we found  $\sim 90\%$  of them to have counterparts in the reference image; in fact almost all sources in the reference image brighter than  $m_R = 18$  have been detected multiple times in the difference images (Fig. 6). This may be due to some of the bright sources being genuinely variable, and others occasionally producing artifacts due to imperfect PSF-matching in the DI pipeline. Only a mere  $\sim 450$  candidates remained without counterparts on the reference image. These sources are candidates for transient sources (different from variable sources in this context). Thus we achieved an almost three orders of magnitude reduction from the initial number of raw DI pipeline detections of  $\sim 2.5 \cdot 10^5$ .

The method presented here is, in principle, well-suited for archival, but not real-time time-domain data analysis. However, for the modern time-domain surveys (for example iPTF and the upcoming LSST), cadences for particular fields could range from hours to even minutes per observation, such that the total number of visits during the night can sum up to more than a few. In such cases, our method can be straightforwardly implemented to search for variable and transient sources near real-time.

The prospects of this method are attractive when it comes to investigating the *populations* of variable and transient objects. It can successfully probe the variable and transient sources even in the regions severely affected by low quality of the DI, for example in the bulge of M31 or in the Galactic plane, which could not practically be achieved earlier in large-scale surveys. The method can be implemented within the automated DI pipeline of the surveys, which will then directly yield outputs ready for astrophysical studies.

We have illustrated the use of the proposed method by applying it to a systematic search for novae in the iPTF data collected during the period from September 2013 to January 2014. To this end, we developed a lightcurve-based filtering procedure to search for novae among the candidates identified by WAVDETECT. We found 15 nova candidates, of which 12 are known novae listed in the MPE nova catalog (Pietsch et al. 2007). Of the remaining 3 candidates, we consider one (candidate 1626, Fig. 8) as a very likely nova based on the shape of its lightcurve and two more (candidates 1687 and 3182) occurred near the end of the analysed data and need further investigation. We have recovered all novae that were reported to occur during the period of these observations in the MPE catalog, except for M31N 2014-01c that erupted at the very end of the iPTF observations. Four (including all three newly discovered candidates) of the novae detected by iPTF (IDs 981, 1626, 1687 and 3182, Fig. 8) occurred far away from the bulge region and thus clearly in the disc of the galaxy. The remaining 11 appear to be concentrated around the bulge region. We do not make any inferences regarding the specific nova rates in the disc and bulge of the galaxy. This will be addressed in the future work based on the  $\gtrsim 10$ -fold larger full iPTF data set.

The fact that all the procedures involved in the novae search — from DI to final lightcurve filtering — are automated will allow one to compute the incompleteness of the sample, as will be done in a future work based on the full iPTF data set. However, we can make a crude estimate of the completeness of the present sample. From Soraisam et al. (2016), the approximate total nova rate for M31 galaxy, counting fast novae, is  $\approx 106/\text{yr}$ . The iPTF M31 field analyzed here encloses roughly 40% of the total stellar mass of M31, and therefore we expect  $\sim 17$  novae to occur during the five-month period of the iPTF M31 observations used here. In our analysis, we have detected 13 (15, if counting the two candidates) novae, which implies a completeness of the sample of  $\sim 76\%$  ( $\sim 88\%$ ). With the high cadence of the iPTF survey, we anticipate the future iPTF M31 nova sample to be particularly useful to constrain population of novae with short decline times. These fast novae are predicted by theories (e.g., Yaron et al. 2005) to occur on massive white dwarfs (WDs) and therefore are imperative for determining the WD mass distribution in novae toward the interesting massive end (see Soraisam et al. 2016).

These results thus demonstrate that the method we have presented in this paper provides an efficient, yet simple, way to analyze the outputs of time-domain data pipelines. By suitably modifying the lightcurve selection algorithm, the method can be easily applied to any survey of any field for probing a wide range of transient sources, including periodic variables such as Cepheids.

## Acknowledgments

MDS is grateful to Niels Oppermann (CITA) for helpful discussions on the Bayesian treatment of obtaining the probability distribution of false matches. MDS thanks the California Institute of Technology, where a portion of this work was completed, for its hospitality. MDS was supported, in part, by the GROWTH internship program funded by the National Science Foundation under Grant No 1545949. AWS is grateful to the NSF for financial support through grant AST-1009566. MG acknowledges partial support by Russian Scientific Foundation (RSF), project 14-22-00271.



## References

- Alard, C. 2000, *A&AS*, 144, 363  
 Alard, C. & Lupton, R. H. 1998, *ApJ*, 503, 325  
 Alcock, C., Allsman, R. A., Alves, D., et al. 1999, *ApJ*, 521, 602  
 Bigley, A., Fuller, K., Hayakawa, K., et al. 2013, *Central Bureau Electronic Telegrams*, 3673  
 Bloom, J. S., Richards, J. W., Nugent, P. E., et al. 2012, *PASP*, 124, 1175  
 Bond, I. A., Abe, F., Dodd, R. J., et al. 2001, *MNRAS*, 327, 868  
 Fabrika, S., Barsukova, E. A., Valeev, A. F., et al. 2013a, *The Astronomer's Telegram*, 5543  
 Fabrika, S., Barsukova, E. A., Valeev, A. F., et al. 2014a, *The Astronomer's Telegram*, 5745  
 Fabrika, S., Barsukova, E. A., Valeev, A. F., et al. 2014b, *The Astronomer's Telegram*, 5745  
 Fabrika, S., Barsukova, E. A., Valeev, A. F., et al. 2014c, *The Astronomer's Telegram*, 5754  
 Fabrika, S., Barsukova, E. A., Valeev, A. F., et al. 2013b, *The Astronomer's Telegram*, 5554  
 Freeman, P. E., Kashyap, V., Rosner, R., & Lamb, D. Q. 2002, *ApJS*, 138, 185  
 Fruscione, A., McDowell, J. C., Allen, G. E., et al. 2006, in *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, Vol. 6270, *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, 62701V  
 Gal-Yam, A., Maoz, D., Guhathakurta, P., & Filippenko, A. V. 2008, *ApJ*, 680, 550  
 Hornoch, K., Manilla-Robles, A., Tudor, V., Vaduvescu, O., & Ramsay, G. 2013, *The Astronomer's Telegram*, 5503  
 Hornoch, K. & Vrástil, J. 2013, *The Astronomer's Telegram*, 5450  
 Kerins, E., Darnley, M. J., Duke, J. P., et al. 2010, *MNRAS*, 409, 247  
 Law, N. M., Kulkarni, S. R., Dekany, R. G., et al. 2009, *PASP*, 121, 1395  
 Masci, F., Laher, R., Rebbapragada, U., et al. 2016, *ArXiv e-prints*  
 Ofek, E. O., Laher, R., Law, N., et al. 2012, *PASP*, 124, 62  
 Pietsch, W., Haberl, F., Sala, G., et al. 2007, *A&A*, 465, 375  
 Rau, A., Kulkarni, S. R., Law, N. M., et al. 2009, *PASP*, 121, 1334  
 Shafter, A. W., Darnley, M. J., Hornoch, K., et al. 2011, *ApJ*, 734, 12  
 Soraisam, M. D., Gilfanov, M., Wolf, W. M., & Bildsten, L. 2016, *MNRAS*, 455, 668  
 Stetson, P. B. 1987, *PASP*, 99, 191  
 Stetson, P. B. 1990, *PASP*, 102, 932  
 Sturm, R., Hofmann, F., Pietsch, W., & Greiner, J. 2013, *The Astronomer's Telegram*, 5384  
 Tang, S., Cao, Y., & Kasliwal, M. M. 2014, *The Astronomer's Telegram*, 5852  
 Tomaney, A. B. & Crotts, A. P. S. 1996, *AJ*, 112, 2872  
 Warner, B. 1995, *Cambridge Astrophysics Series*, 28  
 Williams, R. E. 1992, *AJ*, 104, 725  
 Williams, S. C., Darnley, M. J., Bode, M. F., Keen, A., & Shafter, A. W. 2014, *ApJS*, 213, 10  
 Wozniak, P. R. 2000, *Acta Astron.*, 50, 421  
 Yaron, O., Prialnik, D., Shara, M. M., & Kovetz, A. 2005, *ApJ*, 623, 398  
 Zackay, B., Ofek, E. O., & Gal-Yam, A. 2016, *ArXiv e-prints*

## Appendix A: Probability distribution of false matches

The unknown numbers of true ( $N_t$ ) and false ( $N_f$ ) matches are related to the known number of matched WAVDETECT sources  $N_m$  by the obvious relation

$$N_m = N_t + N_f. \quad (\text{A.1})$$

Assuming a uniform distribution of the sources in the reference image across the region of interest, the expected number of reference image sources occurring randomly within the matching area is given by

$$\mu(N_t) = \pi r^2 (N_r - N_t) / A, \quad (\text{A.2})$$

where  $A$  is the known area under study,  $r$  is the known matching radius and  $N_r$  is the known number of reference image sources in this area. The probability of having one or more random matches for a given source is then

$$p = p(N_t) = 1 - e^{-\mu(N_t)}. \quad (\text{A.3})$$

If  $N_t$  is known, the probability distribution of false matches for the  $N_w - N_t$  sources ( $N_w$  is the known number of total WAVDETECT sources), which is binomial, can be written as

$$P(N_f|N_t) = \binom{N_w - N_t}{N_f} [p(N_t)]^{N_f} [1 - p(N_t)]^{N_w - N_t - N_f}, \quad (\text{A.4})$$

where  $p(N_t)$  is given by Eq. (A.3).

We then relate the above probability distribution to the one that we are interested in, that is  $P(N_f|N_m)$ , by using Bayes' Theorem and marginalizing over various quantities, as

$$P(N_f|N_m) = \sum_{N_t=0}^{\infty} P(N_f, N_t|N_m) \quad (\text{A.5})$$

$$= \sum_{N_t=0}^{\infty} P(N_f|N_t, N_m) P(N_t|N_m) \quad (\text{A.6})$$

$$= \sum_{N_t=0}^{\infty} P(N_f|N_t, N_m) \frac{P(N_m|N_t) P(N_t)}{P(N_m)}. \quad (\text{A.7})$$

Here,  $P(N_m)$  is a normalization constant. Assuming the prior  $P(N_t)$  to be uniform, we introduce the constant  $C = P(N_t)/P(N_m)$ . Then,

$$P(N_f|N_m) = C \sum_{N_t=0}^{\infty} P(N_f|N_t, N_m) P(N_m|N_t). \quad (\text{A.8})$$

The last term in the above equation can be written as

$$P(N_m|N_t) = \sum_{N_f=0}^{\infty} P(N_m, N_f|N_t) \quad (\text{A.9})$$

$$= \sum_{N_f=0}^{\infty} P(N_m|N_f, N_t) P(N_f|N_t) \quad (\text{A.10})$$

$$= \sum_{N_f=0}^{\infty} \delta(N_m - (N_t + N_f)) P(N_f|N_t) \quad (\text{A.11})$$

$$= P(N_f = N_m - N_t|N_t). \quad (\text{A.12})$$

Using Eq. (A.4), it follows

$$P(N_f|N_m) = C \sum_{N_t=0}^{\infty} P(N_f|N_t, N_m) \binom{N_w - N_t}{N_m - N_t} [p(N_t)]^{N_m - N_t} [1 - p(N_t)]^{N_w - N_m} \quad (\text{A.13})$$

$$= C \sum_{N_t=0}^{\infty} \delta(N_m - (N_t + N_f)) \binom{N_w - N_t}{N_m - N_t} [p(N_t)]^{N_m - N_t} [1 - p(N_t)]^{N_w - N_m} \quad (\text{A.14})$$

$$= C \binom{N_w - (N_m - N_f)}{N_f} [p(N_m - N_f)]^{N_f} [1 - p(N_m - N_f)]^{N_w - N_m}. \quad (\text{A.15})$$

Thus, the probability distribution of false matches obtained in Eq. (A.15) appears identical to Eq. (A.4), but with  $N_t$  replaced by  $N_m - N_f$  (Eq. A.1) and a normalization constant  $C$  that can be directly calculated.

It is then straightforward to calculate the expected number of false matches as

$$\langle N_f \rangle = \sum_{N_f=0}^{N_f=N_m} N_f P(N_f|N_m). \quad (\text{A.16})$$

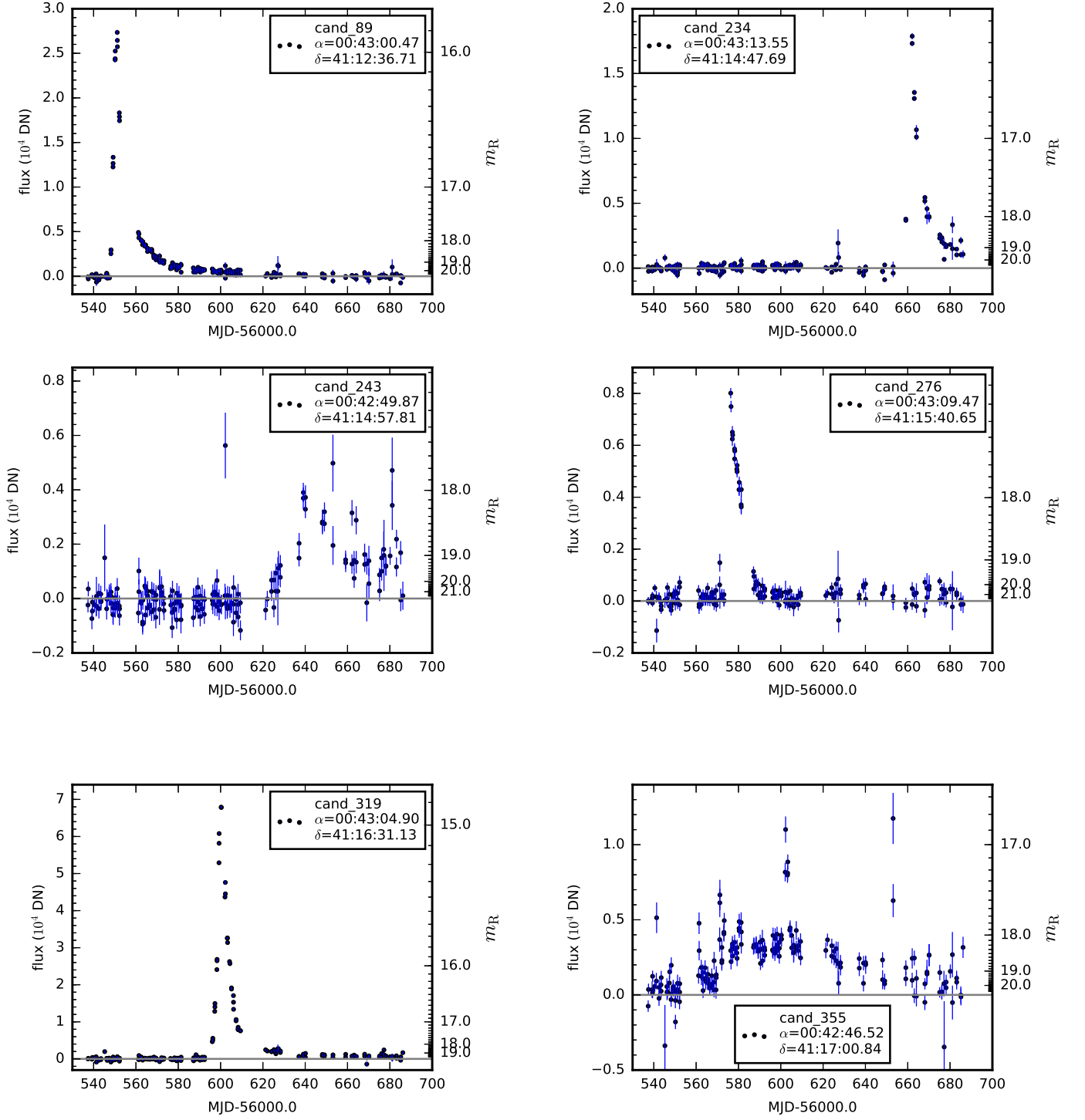


Fig. 8: Lightcurves of the nova candidates obtained from the nova selection algorithm (see Sect. 5.1). The vertical axis on the right side shows the calibrated *R*-band magnitudes corresponding to the flux values on the left vertical axis.

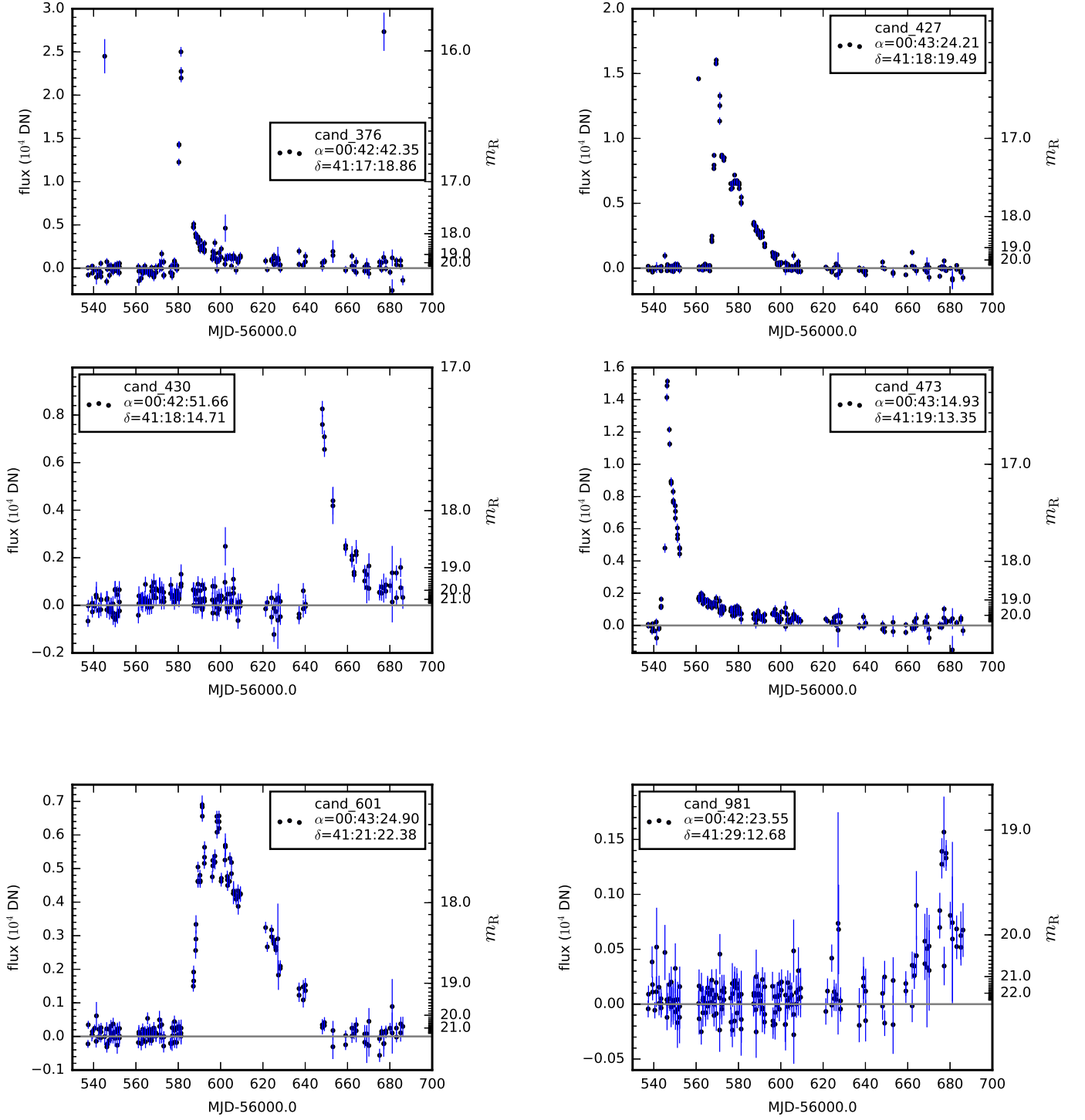


Fig. 8: continued.

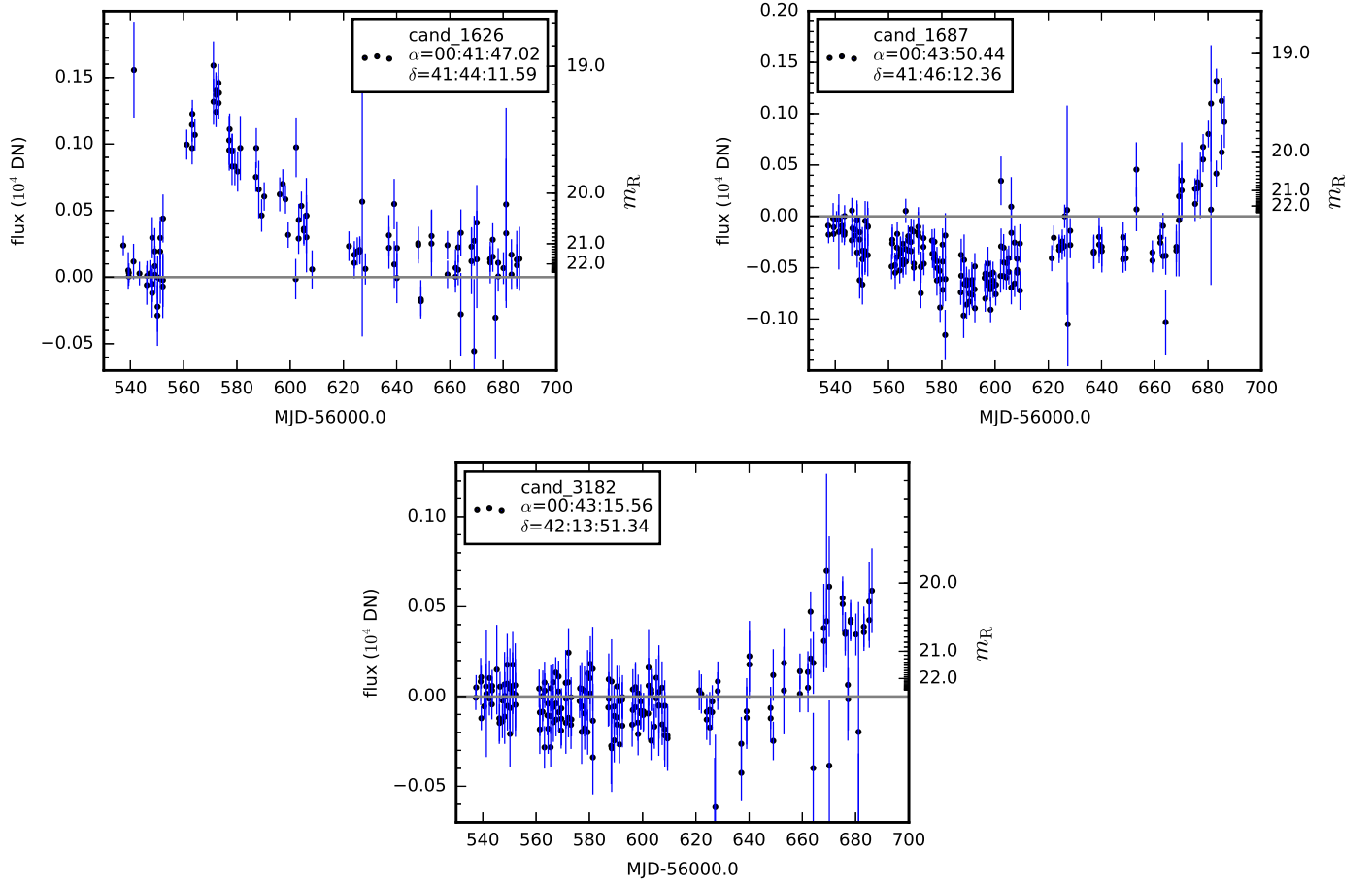


Fig. 8: continued.